Homework # 9

1. Attached you will find a dataset `dataset.csv` containing $n = 1000$ samples (x_i, y_i) with $x_i \in \mathbb{R}^2$ and $y_i \in \{0, 1\}$.

   (a) Visualize the dataset by generating a scatter plot of the x_i and using the value of y_i to color the points.

   (b) Apply a logistic regression to the datasets. Recall, we assume
   $$P(y = 1 \mid x, \alpha) = \frac{1}{1 + \exp[-\alpha \cdot x]} \tag{1}$$
   and then find α that maximizes the log-likelihood. As in a previous homework, add a leading 1 to each x_i so that your fit of the logistic regression gives an optimal $\alpha \in \mathbb{R}^3$.

   (c) Use the fit from (b) to generate a classifier $\Phi(x) : \mathbb{R}^2 \to \{0, 1\}$. Letting α^* be the optimal α from (b),
   $$\Phi(x) = \begin{cases} 1 & \text{if } P(y = 1 \mid x, \alpha^*) \geq .5 \\ 0 & \text{if } P(y = 1 \mid x, \alpha^*) < .5 \end{cases} \tag{2}$$
   Plot the decision boundary of your classifier. The decision boundary will be a line that separate points $\Phi(x) = 1$ from $\Phi(x) = 0$. Add the data points x_i on top. What is the accuracy of your classifier?

2. Consider the polynomial reproducing kernel $k(x, y) = (x^T y + 1)^d$ for $d \in \mathbb{N}$ and $x, y \in \mathbb{R}^2$. Let H be the function space generated by $k(x, y)$.
   $$H = \text{span}\{f(x) : \mathbb{R}^2 \to \mathbb{R} \mid f(x) = k(x, y) \text{ for some } y \in \mathbb{R}^2\}. \tag{3}$$
   Show that H is the space of all degree d polynomials. Write down a basis for H. What is the dimension of H?

3. Continuing with the notation of problems 1 and 2, let
   $$z_i = k(x, x_i) \tag{4}$$
   and consider the dataset (z_i, y_i) for $i = 1, 2, 3, \ldots, n$. (Note $z_i \in H$.) Define the logistic regression model on H by,
   $$P(y = 1 \mid z, \alpha) = \frac{1}{1 + \exp[- < \alpha, z >]}, \tag{5}$$

for $\alpha \in H$. $<,>$ is the inner product of H defined by k. To fit the logistic regression, we need to solve the following optimization.

$$\max_{\alpha \in H} \sum_{i=1}^{n} y_i \log P(y_i \mid z_i, \alpha) + (1 - y_i) \log(1 - P(y_i \mid z_i, \alpha)). \quad (6)$$

Let α^* solve (6).

(a) Show that we can assume $\alpha^* = \sum_{i=1}^{n} \beta_i z_i$.

(b) Rewrite (6) as an optimization in terms of β. Letting $K_{ij} = k(x_i, x_j)$ show that all you need to determine β are the y_i and the matrix K. Using either steepest ascent or Newton's method, find the optimal β for $d = 5$ and $d = 50$.

(c) Let β^* be the optimal β you found in (b). Given β^*, define the classifier $\Psi(z) : H \to \{0, 1\}$ analogous to $\Phi(x)$ in problem 1. We can use Ψ to define a classifier $\tilde{\Psi} : \mathbb{R}^2 \to \mathbb{R}$, For $w \in \mathbb{R}^2$,

$$\tilde{\Psi}(w) = \Psi(k(x, w)). \quad (7)$$

Compute the decision boundary of $\tilde{\Psi}$. One simple way to do this is to put a grid down on $\mathbb{R}^2$. At the grid points evaluate $\tilde{\Psi}$. This will give you a general idea of the decision boundary. Show the boundary for the case $d = 5$ and $d = 50$.